International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Effective Feature Selection and Soft Voting Classifier based Diabetes Detection Using Machine Learning Approaches

Mrs. Uma HR, Asst. professor,
Dept. of Computer Science and Engineering,
BGS Institute of Technology,
BG Nagara, Karnataka

Rohan Gowda R,
4th Year, 8th semesterComputer Science and Engineering,
BGS Institute of Technology,
BG Nagara, Karnataka

***Abstract*—An unnecessarily increased blood glucose level is a symptom of the metabolic disorder diabetes mellitus (DM). Early diabetes detection is reduce the risk factor of kidney disease,heart failure, liver cirrhosis, it might be hazardous. Our motivation is to developed a machine learning models to forecast diabetes in the future. We applied efficient data cleaning, label encoding, normalization and pre-processing methods for better models accuracy. Prediction with synthetic minority over-sampling (SMOTE) implies the equilibrium of an unstable dataset. Different fea-ture selection techniques such as Chi-square test, Information gain attribute evaluator, Extra trees classifier, SelectKbest, and correlation-based approach to finding the key variables. Machine learning algorithms including Random Forest(RF), K-nearest neighbor(KNN), Naive Bayes (NB), Support Vector Machine (SVM), Gradient Boosting (GB), Decision Tree (DT), and Logistic Regression (LR), have been applied in the initial phase. Random Forest accuracy of 95.03% after normalizing the dataset with synthetic minority over-sampling (SMOTE) techniques. We using a voting classifier for ensemble all machine learning models in the second phase. The models achieved soft voting highest accuracy 96.69%. The findings of our models comparing with some current research indicate that it can offer greater accuracy. Now, we can accurately forecast diabetes.**

***Index Terms*—Diabetes, SMOTE Oversampling, Ensemble approach, Soft voting, Machine learning.**

I. INTRODUCTION

Diabetes mellitus is a significant social, health, and eco- nomic problem that is rapidly becoming a global one. A metabolic disorder unnecessarily increased blood glucose lev- els is known as diabetes mellitus (DM) [1]. Due to decreased insulin production and elevated blood sugar levels, a person's metabolism is significantly impacted. Body cells react to an unfavorable way when there is less insulin in the circulation [2]. There are three different forms of diabetes as well as pre- diabetes condition. When pancreatic cells are unable to make enough insulin, outside sources must be used to administer insulin injections in order to keep the body's glucose levels stable is refers to as Diabetes type 1. This kind of diabetes typi- cally affects younger people (less than 20) and a family history of diabetes [3]. When the metabolic system unable to process the meal, type 2 (Adult onset) diabetes develops, which raises blood sugar levels. Inactive lifestyle, weakened immunological and neurological systems are common factors. Most people with type 2 diabetes are between the ages of 45 years and 60 years. Pre-diabetes also known as borderline diabetes, elevated sugar level in the blood that are not high enough to be diagnosed as diabetes [4]. The WHO estimates 470 million diabetics in the globe as of 2019 and the number will increase to 700 million by 2045. Diabetes can be controlled and a person's life can be saved with early diagnosis [5]. Individuals who may have diabetes must undergo a variety of tests and procedures to accurately diagnose the diabetics. These could be duplicate or unneeded medical procedures, which resultsin complication as well as time and resource inefficiency [6]. This research investigates the possibility of predictingdiabetes by using several characteristics associated with the condition. With the help of the clinical Diabetes Dataset, we can predict diabetes using machine learning algorithms and ensemble techniques.

The following are the study's main goals.

- First, we proposed a two-phase framework and demon- strated its accuracy in predicting diabetes.
- Second, we used appropriate pre-processing techniques, SMOTE for data balance, and a variety of feature selec- tion techniques to isolate the top traits from the group.
- Finally, we used a variety of machine learning algorithms and ensemble learning techniques, such as soft voting classifier, to improve the performance of ML models. The model's performance was then compared to earlier research.

The remaining portion of the work are structured as follows: The interrelated works are described in Section II. The specie- fictions of the suggested machine learning models presented in Section III. The results are displayed in Section IV. The paper's conclusion and future research offers in Section V.

II. LITERATURE SURVEY

Numerous techniques for detecting and preventing diabetes have been discovered in recent years. This section covers significant machine learning research that is related to this research.

jackins et al. [7] developed a technique for clinical data analysis which determined the correlation between the at- tributes. One features can eliminated other and used for classi- fication. In order to predict the disease, the authors calculated the connection between features after creating a confusion matrix. The authors used NB, RF classification algorithm for data analysis and comparison.

In [8] the authors proposed a method that optimized the feature correlation to choose key feature. The classification carried using the fast miner tool, which calculated the core- lation between the attributes. They applied machine learning algorithm SVM, NB, RF DT, KNN and got accuracy 98.20%, 98%, 82.30% respectively.

Pranto et al. [9] predicted diabetics mellitus using several Machine learning algorithm such as RF, KNN, DT, NB. They collected dataset from Kaggle and Kurmitola General Hospital diabetes dataset. They examine the correlation matrix for their dataset. They got the highest accuracy rate of 77.90% forRF, 81.20% for KNN, 79.50% for DT and 79.90% for NB respectively.

In [10] authors proposed a machine learning models for dia- betics risk factor for North Kashmir. They offered an approach employing K-fold, extraction of features, data processing, SMOTE for data balancing. They were used RF, SVM, GBC, DT, MLP and LR. The random forest accuracy was 98% which was greater than other algorithm's. The overall accuracy was 90.99%, 92%, 97%, 96%, 69% for MLP, SVM, GBC, DT, LR
respectively.

In [11] the authors provided an approach in which the PIDD was subjected to machine learning techniques, including NB, SVM, and DT. This approach is a relatively straightforward for classification methods. The authors perform precision, recall, ROC and accuracy. Finally got greater accuracy NB for76.30%.

Liza et al. [12] suggested ML methods on two datasets that are openly accessible and evaluated. The models used LR, KNN, AdaBoost and MLP that accurately identified the early diabetics. The authors employed StackingCV classifier to ensemble these four models. Finally MLP performed the highest accuracy of 91% and 93% for two datasets.

In [13] the authors used machine learning algorithms LR, XGBoost, GB, DT, Extra Trees, RF, Light gradient boosting machine for type-2 diabetics. The study used PIMA dataset for predicting DM. The highest accuracy of LGBM perform 95.20%.

In [14] authors developed soft voting classifier to ensamble three machine learning algorithm RF, LR and NB. They collected PIMA dataset and used several ML algoriths such chine Learning dataset repository, which is a free and open- source platform [15].The dataset includes on 952 people, whom 266 had diabetes and 686 non diabetes patients. The dataset had 18 variables, where one is target variable for prediction and 17 contains patient's information. **Table I** lists the dataset properties in details.

TABLE I: Features Of The Dataset

Attributes Name	Data Type	Description	Value Range
Age	Object	Age in Year	18 or above
Gender	Object	Gender of the participant	Male/ Female
Family Diabetes	Object	Family history with diabetes	Yes or No
highBP	Object	Diagnosed with high blood pressure	Yes or No
Physically Active	Object	Walk/run/ physically active	Less than half an hour One hour
BMI	Float	Body Mass Index	Numeric
Smoking	Object	Habit of Smoking	Yes or No
Alcohol	Object	Alcohol consumption	Yes or No
Sleep	Integer	Hours of Sleep	Numeric
SoundSleep	Integer	Sound Sleep Hours	Numeric
Regular Medicine	Object	Regular intake of medicine	Yes or No
JunkFood	Object	Junk food consumption	Yes or No
Stress	Object	Level of stress	Not Some times /Often / Always
BPLevel	Object	Blood pressure level	High/normal/low
Pregancies	Float	Number of pregnancies	Numeric
Pdiabetes	Object	Gestation diabetes	Yes or No
Uriation Freq	Object	Frequency of urination	Not much/ Quite much
Diabetic	Object	Class or target	Yes or No
Total participants = 952, Total variables = 18			

as LR, KNN, SVM, NB, DT, RF, adaBoost, Bagging, GB,XGB, CatBoost and soft voting calssifier. They got accuracy

79.08%, 97.27% in soft voting than other algorithms.

III. METHODOLOGY

A. *Dataset and Features Information*

We conducted a diabetes predictive single dataset. The dataset obtained from University of California, Irvine, Ma-

B. *Data Pre-processing*

- **Cleaning Data:** The dataset information was retrieved into its raw form. The dataset has been cleaned by doing a number of tasks, such as removing outliers, handling missing values, data normalization, encoding and so forth. For managing missing values, the mean value of every single attribute was implemented to enhance the performance of the model. Converting textual categorized value into a numeric value known as label encoding. Data needs to be initially normalized for the range of attributes, which was done using min-max normalizing approach. To eliminate outliers we applied DBSCAN algorithm [5].Min–max normalization equation is:

$$X(norm) = \frac{X - min(x)}{max(x) - min(x)} \quad (1)$$

Where X is feature's value, min(x) is minimum value, and max(x) is maximum value.

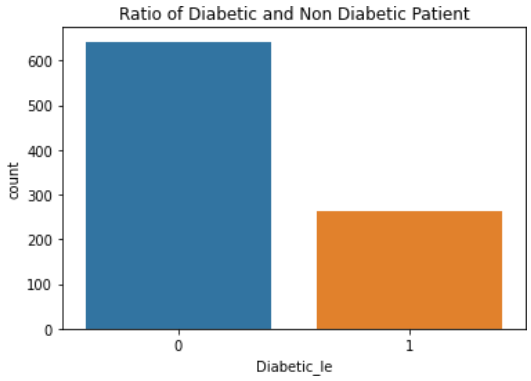
- **Data Balancing Using SMOTE:** Data balancing involves rebalancing data from unbalanced data. Data balanced strategy applied to enhance the amount of irregularly distributed data and minimize the number of records

to avoid overflowing. The SMOTE technique generates new minority class instances to attain a balance over the minority and majority category sizes in order to address the imbalanced dataset problem [10]. These exampleswere developed using variables from the preliminary dataset in an effort to closely resemble real minority class instances [16]. To enhance minorities in the class, SMOTE Equation given is below:

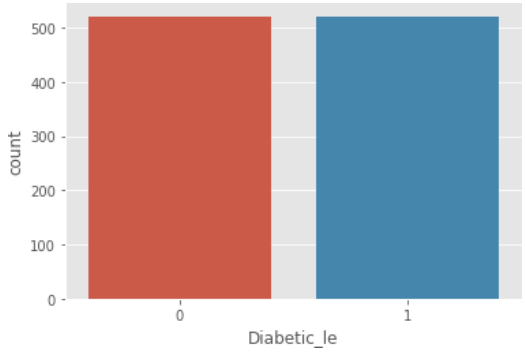
$$X_{syn} = X_i + (X_{knn} - X_i) \times t \quad (2)$$

Where Xi is feature vector and Xknn identifies the K- nearest neighbors. Then it determines how the feature vector and k- nearest neighbor differ from one another. To find feature vectors, it repeats the previous processes.

After data cleaning reduces the duplicate values, restoring missing value using mean filter approach and encoding data values into a numerical format. The dataset includes informa- tion on 905 patients, whom 263 had diabetes and 642 non diabetic patients. **Fig.1a** shows before sampling, which has diabetic vs normal (0 indicates normal and 1 indicates diabetic) and the value after sampling shown in **Fig.1b**.



(a) Before SMOTE



(b) After SMOTE

Fig. 1: Ratio of diabetics and non-diabetics patients

C. *Feature Selection*

We used chi-square test, information gain attribute evaluator, extra trees classifier and correlation-based feature selection approach. We used select best feature (SelectKbest) from all methods based on scores of training.

- **Chi-Square Test:** The chi-square test was performed to determine whether the class mark were independent or not in a feature [17]. We calculate top feature applied chi- square feature technique. The effective chi-square values analysis with a target variable. The chi-square as follow:

$$X^2 = \frac{\sum (O_i - E_i)^2}{E_i} \quad (3)$$

where, Oi be the observed value and Ei enhance expected value.

- **Information Gain Attribute Evaluator:** The quantity of information about the class measured by the information gain attribute evaluator. We got no information from features that are not related to one another. Features observed based on high information gain efficiency [1]. Information gain eliminating random attribute informa- tion to optimized model. We calculated IG as follow:

$$E(S) = \sum_{i=1}^{C} -P_i log2 P_i \quad (4)$$

P represents the proportion of cases that fall under the class.

- **Extra Trees classifier**: It generates a large number of decision trees without pruning and combining the pre- diction of decision trees for majority vote classification. A random sample of K features given to each tree, which chooses the best feature.
- **SelectKBest:** The SelectKBest method was chosen based on the attributes highest score. We applied algorithm for both classification and regression data by adjusting the ”scoring function” parameter. It determines the relation- ship between two categorical variables that reflects the dataset. It facilitates the removal of irrelevant data and shortens the training period.
- **Correlation Based Feature Selection:** The Pearson’s correlation between the attribute and the class used to determine the value of the attribute. By using a weighted average, one can find the overall correlation for a nominalproperty [12].

The visualization result after using all feature selection techniques given in **Fig.2**.

Features chart using correlation based feature selection shown in **Fig.3** and **Table II** highlights the features that we have selected.

TABLE II: Selected Top Features

Selected best Features	
BMI	BPLevel le
Age le	highBP le
Pregancies	Sleep
RegularMedicine le	Stress le
FamilyDiabetes le	SoundSleep
PhysicallyActive le	Diabetic le

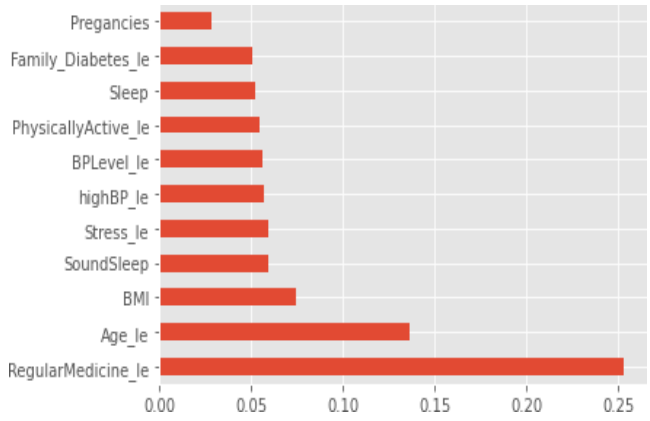


Fig. 2: Various top features based on scores

We have selected **eleven** effective features based on scores. The selected features are common in all of the algorithms such as chi-square test, information gain attribute and extra trees classifier. The confusion arise that one feature utilized higher scores in extra trees classifier method. The Chi-square test and Information gain evaluator shows less scores for that feature. In correlation attribute some features shows less value where others have higher values. To solve this problem, we applied SelectKbest feature selection method to select best features among all four(Chi-square, IGA, ET, Correlation) techniques.

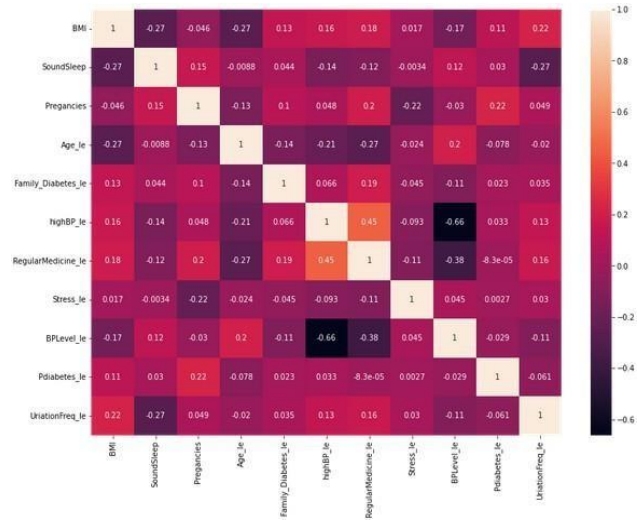


Fig. 3: Correlation analysis of best features

D. Proposed Models

We have implemented a machine learning categorization models for assessing risk of persons diabetes at a preliminary phase. We suggested two phase models selection procedure for ML approaches to forecast diabetes. Then preprocessed dataset by cleaning, data encoding, filling missing values, and utilizing SMOTE for data balancing. We used DBSCAN filter for noisy data, handle outliers to select the best model. Selecting top fea- tures using a combination of feature selection approaches (Chi- square, IGA, ET, correlation) to obtain the relevant variables. We used another feature method SelectKbest to select best features among them. In data splitting 80% of the preprocessed data were utilized to build the prediction model, while the remaining 20% were used for testing. The machine learning algorithms (RF, KNN, NB, SVM, DT, LR, and GBC) used to evaluate accuracy and classification performance. afterwards, voting classifier ensemble learning approach applied to im- prove the results of machine learning model. **Fig.4** illustrates the block diagram of our suggested approach for detecting diabetes.

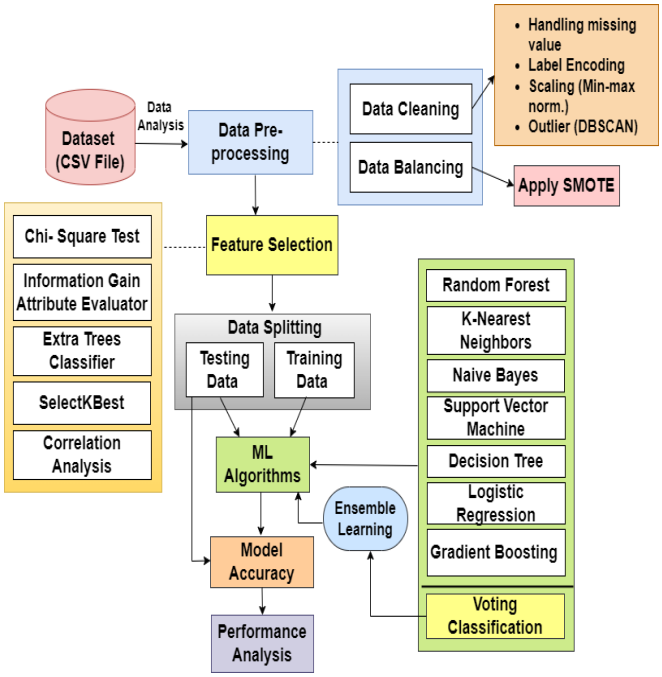


Fig. 4: Proposed framework for diabetes detection

E. Classification of Algorithms

- **Random Forest (RF):** In order to provide one result, the RF combines the output values or results of various Decision Trees. The DT under consideration used as a foundational row sampling method and column sampling method. To increase accuracy, the number of training sets need to increased while the variable also minimized depending on the inputs [2].
- **Naive Bayes (NB):** The Naive Bayes Classifier is one of the simplest and finest classification algorithms for creating short machine learning models that can predict outcomes quickly. For classification, it is one of the most effective ML algorithms. The Bayes theorem is extended in that each feature is assumed to be independent. Itis used for many different things, Predicting and text classification [3].

- **K-Nearest Neighbor (KNN):** The KNN is one of the mostly used supervised ML algorithm. KNN is used in a regression and classification model. The KNN considers the distance between testing data and the training data. The category that best fits the new case's parameters is then chosen from thoseoffered [8].
- **Support Vector Machine (SVM):** A number of SVM classes or objects are used to construct the higher dimen- sional space. The SVM and class corner points used to calculate the average between the classes starting from the hyperplane's centre point. The kernel is the mostimportant part of the SVC. These kernels have been modified to take into account the type of data they received [10].
- **Decision Tree (DT):** Decision Trees are one of the most reliable used algorithms for predicting and classifyingdata.An attribute test is represented as an internal nodein a DT design. The leaf nodes (terminal nodes) having the appropriate class labels serve as representations of theresults of the corresponding tests [11].
- **Logistic Regression (LR):** Logistic regression ML clas- sification predict the probability of a target variable and vote. LR not employed in less dataset than features this results in overfitting. We used LR for binary classifica- tion. In order for an output belong to either of the two classifications, we categorize it (1 or 0). In ML, we utilizesigmoid to convert predictions to probabilities [13].
- **Gradient Boosting Classifier (GBC):** Another machine learning technique for solving regression and classifica- tion issues is gradient boosting. This prediction model was created by combining many poor predictions, the majority of which were DT. The model is built in phases, just like previous boosting techniques. The model then compressed using an approach to unlimited differentiableloss function optimization [14].
- **Ensemble Learning**: The use of numerous distinct classi- fiers can improve the classification accuracy of the model. When operating on the similar topic, multiple ML algo- rithms collaborate to increase prediction performance.
- **Voting Classifier(VC):** To choose the best from a listof multiple alternatives, use a voting system. As a result, many classifiers can choose from a variety of alternatives.In light of the majority's selections a final option is determined. A better answer can discovered on several algorithms applied to the same issue. Not everyone com- mits the same error while using ensembles in various categories.

IV. RESULT ANALYSIS

Our suggested model used RF, KNN, NB, DT, SVM, LR and GBC machine learning algorithms on DM dataset. Then the model ensemble using voting classifier. The dataset split into 80% and 20% respectively. The evaluation metrics accu- racy, precision, recall, f1-score applied to compare the model performance. Evaluation metrics rely on four classification labels true positive(TP), true negative(TN), false positive(FP)and false negative(FN).

Accuracy: Divide total number of participants sample bythe sum of true positive and true negative results.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (5)$$

Precision: The percentage of all positively expected obser- vations accurately anticipated.

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (6)$$

Recall: The percentage of results that properly anticipated positive observations. It also called sensitivity.

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (7)$$

F1-score: Calculated the precision and recall weighted av- erage and consider both false positive (FP) and false negative(FN).

$$F1 - score = \frac{2* Recall* Precision}{Recall+ Precision} \times 100 \quad (8)$$

After employing the SMOTE approach to equalize the dataset, Random forest perform greater accuracy rate 95.03%. The accuracy results of soft voting was 96.69%, **Fig.5** high- lights the accuracy of various algorithms with VC.

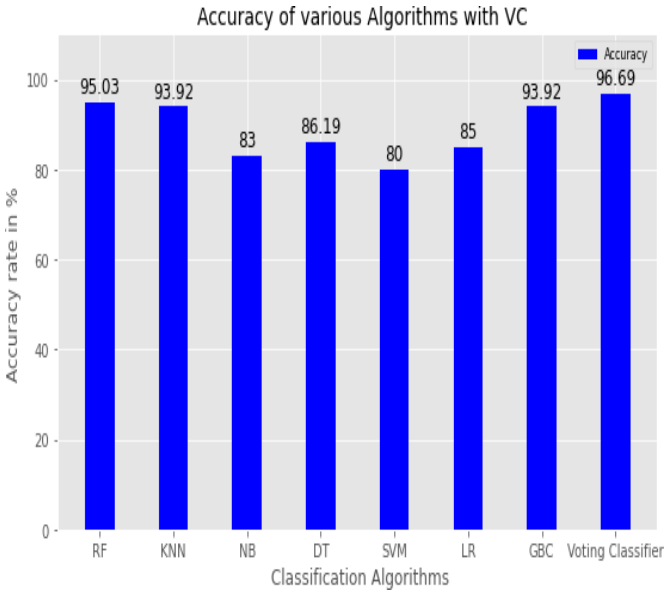


Fig. 5: Accuracy comparison with VC

The overall accuracy, performance matrices for seven al- gorithms with ensemble approaches are given in **Table III** and the total findings of the experiment in terms of accuracy, precision, recall, and f1-score are shown in **Fig.6**. The receiver operating characteristic(ROC) curve displays in **Fig.7** where RF and VC provide better performance.

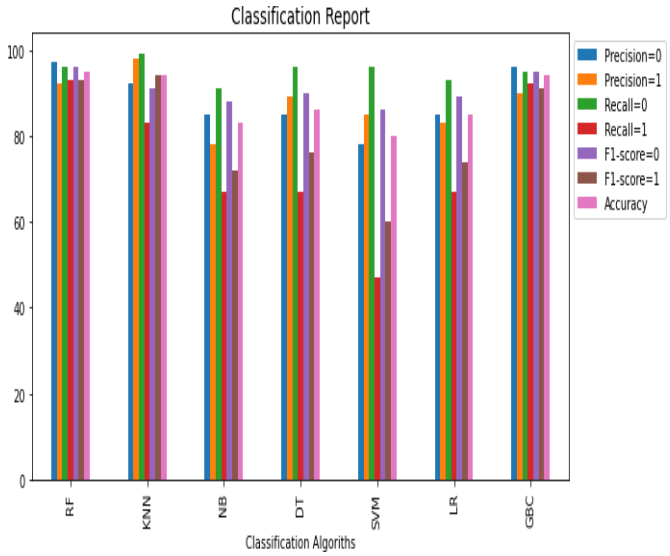


Fig. 6: Overall performance metric

TABLE III: Overall accuracy and other performance metrics

Models	Outcome	Precision	Re-call	F1-score	Accuracy %	Voting (Soft)
RF	0	0.97	0.96	0.96	95.03	96.69
	1	0.92	0.93	0.93		
KNN	0	0.92	0.99	0.96	93.92	
	1	0.98	0.83	0.90		
NB	0	0.85	0.91	0.88	83	
	1	0.78	0.67	0.72		
DT	0	0.85	0.96	0.90	86.19	
	1	0.89	0.67	0.76		
SVM	0	0.78	0.96	0.86	80	
	1	0.85	0.47	0.60		
LR	0	0.85	0.93	0.89	85	
	1	0.83	0.67	0.74		
GBC	0	0.96	0.95	0.95	93.92	
	1	0.90	0.92	0.91		

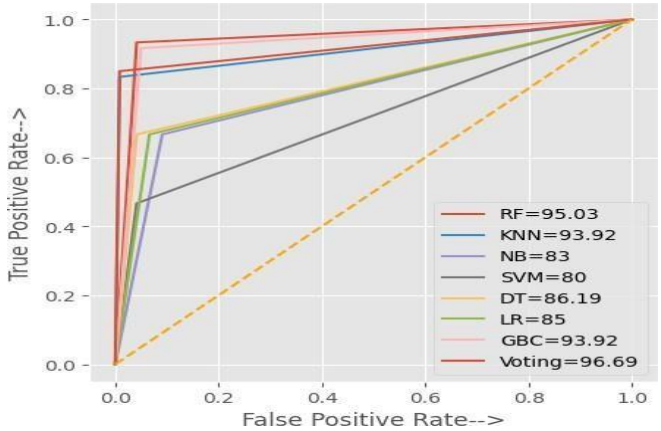


Fig. 7: ROC curve for proposed DM model

- **Comparative Analysis**: In this research, the SMOTEhas been employed for oversampling. When balancing techniques are used on a DM dataset, the accuracy of each classifier changes. **Fig.8** describe the performance accuracy using with SMOTE and without SMOTE.

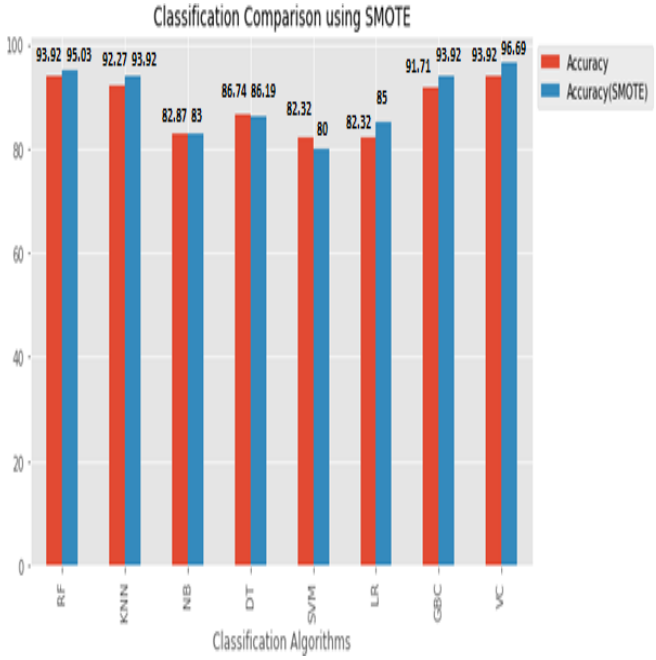


Fig. 8: Comparison SMOTE and Without SMOTE

- **Comparative analysis with other studies**: The authors Tigga and Garg collected same dataset as we used in this study. Machine learning algorithms performed for predicting DM. We used their dataset and discoveredmore accuracy than they have. The accuracy of our pro- posed method on same dataset is compared in **Table IV**. Another performance comparison between Pima indian diabetes dataset and our proposed method on diabetes dataset demonstrated in **Table V**.

TABLE IV: Accuracy Comparison On Same Dataset

Method Name	Accuracy in %	
	Tigga and Garg [18]	Our Proposal
RF	94.10	95.03
KNN	77.30	93.92
NB	80.60	83
DT	84.00	86.19
SVM	86.50	80
LR	85.70	85
GBC	—-	93.92
Ensemble VC	—-	96.69

TABLE V: Comparison with other experiments

Reference	Applying Method	Accuracy (%)
Rubaiat et al. [19]	MLP classifier with feature selection approach	85.15
Ahmad et al. [20]	DT and MLP classifier	81.9
Kumari et al. [14]	Ensemble approach using soft voting classifier	79.08
Pranto et al. [9]	KNN, RF, NB, DT classifier	77.90
Proposed ML Approach	Feature selection with soft voting classifier	96.69

Random Forest and ensemble approach of soft voting clas- sifier performs better than the other existing approaches in the initial diabetes detecting process although ensemble is a time-consuming and laborious procedure.

V. CONCLUSION

Diabetes is an alarming and ongoing disease. Early diabetes detection makes it possible to treat patients more successfully. In this work, various machine learning-based classification models for estimating people's risk of developing diabetes were analyzed. We suggested two-phase model selection pro- cedure for machine learning approaches to forecast diabetes. Before to preprocessing, data were cleaned up and normal- ized. Discovering the crucial variables by selecting the best features using a variety of feature selection approach. From various ML classifier the most reliable classifier is random forest with performance accuracy 95.03%. Ultimately, voting classifier ensemble learning approaches were used to improve the efficacy of machine learning models. The accuracy results of the soft voting is 96.69%. Our suggested model is effective and quicker that doctors would be able to quickly diagnose patients and estimate their probability of getting diabetes. In future, we will go to enhance multiple newly available datasets with machine learning classifiers being employed, which may be improved more correctly to predict disease mellitus.

REFERENCES

[1] Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta, G. C. Sampada, "A Novel Approac h for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods", vol. 2022, Article ID 3820360, 11 pages, 2022. https://doi.org/10.1155/2022/3820360.

[2] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learn- ing Classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[3] Md. Maniruzzaman, Md. J. Rahman, B. Ahammed, and Md. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," Health Inf Sci Syst, vol. 8, no. 1, p. 7, Dec. 2020, doi: 10.1007/s13755-019-0095-z.

[4] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," BMC Med Inform Decis Mak, vol. 19, no. 1, p. 211, Dec. 2019, doi: 10.1186/s12911-019-0918-5.

[5] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," International Journal of Cog- nitive Computing in Engineering, vol. 2, pp. 229–241, Jun. 2021, doi: 10.1016/j.ijcce.2021.12.001.

[6] S. Albahli, "Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes Detection," j med imaging hlth inform, vol. 10, no. 5, pp. 1069–1075, May 2020, doi: 10.1166/jmihi.2020.3000.

[7] Jackins, V., Vimal, S., Kaliappan, M. et al. "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes".J Supercomput 77, 5198–5219 (2021). https://doi.org/10.1007/s11227- 020-03481-x.

[8] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," J Big Data, vol. 6, no. 1,p. 13, Dec. 2019, doi: 10.1186/s40537-019-0175-6.

[9] B. Pranto, Sk. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh," Information, vol. 11, no. 8, p. 374, Jul. 2020, doi: 10.3390/info11080374.

[10] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–12, Oct. 2022, doi: 10.1155/2022/2789760.

[11] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585,2018, doi: 10.1016/j.procs.2018.05.122.

[12] F. R. Liza et al., "An Ensemble Approach of Supervised Learning Algo- rithms and Artificial Neural Network for Early Prediction of Diabetes," in 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, Dec. 2021, pp. 1–6. doi: 10.1109/STI53101.2021.9732413.

[13] B. S. Ahamed, M. S. Arya, and A. O. Nancy V, "Prediction of Type- 2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques," Front. Comput. Sci., vol. 4, p. 835242, May 2022, doi: 10.3389/fcomp.2022.835242.

[14] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[15] UCI Machine Learning Repository. (1998). Diabetes data set. https://archive.ics.uci.edu/ml/datasets/diabetes.

[16] F. Rustam et al., "Predicting pulsar stars using a random tree boosting voting classifier (RTB-VC)," Astronomy and Computing, vol. 32, p. 100404, Jul. 2020, doi: 10.1016/j.ascom.2020.100404.

[17] S. Jahan et al., "Automated invasive cervical cancer disease detection at early stage through suitable machine learning model," SN Appl. Sci., vol. 3, no. 10, p. 806, Oct. 2021, doi: 10.1007/s42452-021-04786-z.

[18] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," Procedia Computer Science, vol. 167, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[19] S. Y. Rubaiat, M. M. Rahman, and Md. K. Hasan, "Important Feature Se- lection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection," in 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, Dec. 2018, pp. 1–6. doi: 10.1109/CIET.2018.8660831.

[20] A. Ahmad, A.Mustapha, E. D. Zahadi, N. Masah, and N. Y. Ya-haya, "Comparison between Neural Networks against Decision Treein Improving Prediction Accuracy for Diabetes Mellitus," in Digital Information Processing and Communications, V. Snasel, J. Platos, and E. El- Qawasmeh, Eds., in Communications in Computer and Information Science, vol. 188. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 537–545. doi: 10.1007/978-3-642-22389-147.